Sören Sonnenburg,† Gunnar Rätsch,‡ Fabio De Bona,‡

† Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany
‡ Friedrich Miescher Laboratory of the Max Planck Society, Spemannstr. 39, 72076 Tübingen, Germany
Machine Learning Toolbox SHOGUN features algorithms:

- to learn 2-class classification and regression problems
- to train hidden markov models
- toolbox’s focus is on kernel methods esp. Support Vector Machines (SVMs) for computational biology
- also implements a number of linear methods like Linear Discriminant Analysis (LDA), Linear Programming Machine (LPM), (Kernel) Perceptrons
Support Vector Machine

• given: points $\mathbf{x}_i \in \mathcal{X}$ ($i = 1, \ldots, N$) with respective labels $y_i \in \{-1, +1\}$

• in training hyperplane that maximizes margin is chosen

Decision function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$
• SVM decision function in kernel feature space:

\[
f(x) = \sum_{i=1}^{N} y_i \alpha_i \Phi(x) \cdot \Phi(x_i) + b
\]

= \kappa(x, x_i) \quad (1)

• find parameters \(\alpha\) by solving quadratic optimization problem
• **Spectrum Kernel**
  - count k-mers in each sequence, Spectrum Kernel is sum of product of counts

• **Weighted Degree Kernel**

\[
k(S_1, S_2) = w_7 + w_1 + w_2 + w_2 + w_3
\]

• **Weighted Degree Kernel with Shifts**

\[
k(x_1, x_2) = w_{6,3} + w_{6,3} + w_{3,4}
\]
Features

- **SHOGUN** interfaces to Matlab™, Octave and Python and R

- provides generic SVM object interfacing to several different SVM implementations, among them the state-of-the-art LibSVM and SVM\textsuperscript{light}

- SVMs can be trained using a variety of common kernels (Linear, Polynomial, Gaussian) and recent String Kernels (TOP, Fisher, Locality Improved, Spectrum, Weighted DegreeKernel (with shifts))

- kernels can be combined; weighting can be learned using Multiple Kernel Learning.

- input feature-objects can be dense, sparse or strings and of type int/short/double/char; can be converted into different feature types.

- multiprocessor parallelization ⇒ able train on 10 million examples

... and many more...
TACKLED BIOINFORMATICS PROBLEMS

- Protein Super Family classification

- Splice Site Prediction (*C. elegans*, *Drosophila*, *Human* etc.)

- Alternative Splice Site Prediction (Exon Skipping, Intron retention, alternative 3’ or 5’ ends)

- Interpreting the SVM Classifier

- Splice Form Prediction (Learn segmentation)

- Promoter Prediction

⇒ very generic
Bioinformatics Demo:

• Position Independent (e.g. Tissue Classification using Promotor Region)

AAACAAAAACGTAAC TAATCTTTTTAGAGAGAACGTTTTCAACCATT TTGGAG
AAGATTAACCTCATCACAGATTTTCATTACATAGATATAATTCAAAAAATT
CACTCCCCAATCAACGATATT TTTAAAAATCACTAACACATCCGTCTGTGC

– Task: separate DNA strings, ‘-’ class random ACGT, ‘+’ class contains ‘AAAAA’ motif

• Position Dependent (e.g. Splice Site Classification)

AAACAAAATAAGTAAC TAATCTTTTAAAGAAGAACGTTTTCAACCATT TTGGAG
AAGATTAAGAAAAAAAACAAATTTTTTAACATTTACAGATATAATAATTCAATT
CACTCCCCAATCAACGATATT TTTAAAAATCACTAACACATCCGTCTGTGC

– Task: separate DNA strings, ‘-’ class random ACGT, ‘+’ class ‘AA’ in the middle

• Mixture Position Dependent/Independent (e.g. Promoter Classification)

AAACAAAATAAGTAAC TAATCTTTTAAAGAAGAACGTTTTCAACCATT TTGGAG
AAGATTAAGAAAAAAAACAAATTTTTTAACATTTACAGATATAATAATTCAATT
CACTCCCCAATCAACGATATT TTTAAAAATCACTAACACATCCGTCTGTGC

– Task: separate DNA strings, ‘-’ class random ‘ACGT’, ‘+’ class ‘AAA’ in the middle shifted ±15
**Generic Demo:**

- **Support Vector Classification**
  - Task: separate 2 clouds of gaussian distributed points in 2D

- **Support Vector Regression**
  - Task: learn a sine function

- **Hidden Markov Model**
  - Task: 3 loaded dice are drawn 1000 times, find out when which dice was drawn
• **SHOGUN** is a large scale machine learning toolbox
  ⇒ able to train on 10 million examples

• unified SVM framework + many string kernels suitable for comp. biology

• Algorithms: HMM, LDA, LPM, Perceptron, SVM, SVR + many kernels

**We need your help:**

• Documentation
• Examples
• Testing
• Test Suite

**Source Code is freely available under the GPLv2.**
http://www.fml.tuebingen.mpg.de/raetsch/projects/shogun